# Preserving State Government Web Publications -- First-Year Experiences

Larry S. Jackson
Graduate School of Library and Information Science, University of Illinois
501 East Daniel, Champaign, IL 61820
lsjackso@uiuc.edu
http://www.isrl.uiuc.edu/pep

## Abstract

Government information publishing on the web encounters differing expectations concerning the permanence of documents. Presumptions of informality in web authoring and publishing widely persist. But, terms like "government documents" convey, at least to the layman, an expectation of formality, official content, and permanence. If permanent, searchable archives of electronic documents are to be constructed, issues in economical automated assembly and cataloging of multiple simultaneous versions of retained materials must be dealt with. High-quality metadata might contribute to these needs, as well as to enhanced search effectiveness now. This paper reports on the first year's operation of an automated web archival facility designed for interoperability with the Find-It! Illinois metadata-aware search engine for the state government documents of Illinois. Comparison statistics for the state government of Arizona are also provided. The Preserving Electronic Publications system was developed from open-source materials, under a National Leadership Grant from the Institute of Museum and Library Services.

## 1. Introduction & Project Background

Government information on the web encounters differing expectations concerning the permanence of documents. Being only about a decade removed from the introduction of the web, informal notions of web authoring and publishing persist in many circles. But, terms like "government documents" convey, at least to the layman, an expectation of formality, official content, and permanence. If permanent archives of electronic documents are to be constructed, issues in economical automated assembly and cataloging of multiple simultaneous versions of retained materials must be dealt with. High-quality metadata might contribute to these needs. Further, issues of document construction standards are very important, both as they affect gathering processes, and as they bear on the long-term accessibility of electronic documents.

Under a National Leadership Grant "Preserving Electronic Publications" (PEP) (PEP, 2001) from the Institute of Museum and Library Services (IMLS), the Illinois State Library (ISL) and the Graduate School of Library and Information Science (GSLIS) at the University of Illinois, Urbana-Champaign, constructed and operated electronic document archives for 118 State of Illinois websites for the past year. These archives were constructed using web "spider" software to automatically traverse all embedded hyperlinks within web pages, and to download all files so referenced. Files were retained using a version control system, to make available any prior version and specific information concerning the differences between versions. PEP was also run, one time, for 84 State of Arizona websites to produce the comparative statistics reported here. The papers and software associated with this portion of the PEP project are available on the web at (Jackson, 2002). The PEP method of periodically obtaining and retaining copies of web documents could be used at nominal cost by other states, by federal agencies, by corporations, or by groups interested in preserving the web portion of their history.

We examined the complete web-accessible electronic document inventories of Illinois and Arizona. PEP archive operation continues, for Illinois, as part of an IMLS Library Services and Technology Act grant to GSLIS. Historically interesting materials are being gathered as our ongoing archival work spans a state election in which the Governor and a substantial majority of the state-level elected officers were replaced. Table 1 lists overall website size, and count information concerning document types and the number of useful HTML META tags encountered. Further analysis was done of all markup-language documents to

determine the nature and extent of metadata incorporation. Generally speaking, metadata authoring at the individual state agency level was verified to be "just beginning".

The contradictory expectations of permanence and formality were exemplified here in a variety of ways. For example, stylistic "look-and-feel" changes were frequently done that rarely would have been done to materials in print media. Or, website host machine names were often renamed, whereas an organization would seldom revise other locators such as telephone numbers or mailing addresses. And, document storage exhibited multiple conventions, ranging from the formal taxonomies and role-based naming through the highly informal and idiosyncratic. A parallel question arises concerning the level of formality and consistency that might be anticipated in locally authored metadata embedded within these documents.

| Measure | Illinois | Arizona |
|---|---|---|
| Number of websites | 118 | 84 |
| Total size of the download, in bytes | 27,671,171,072 | 10,895,245,312 |
| Number of files | 365,151 | 227,960 |
| Number of markup-language files | 232,408<br>63.6% | 156,188<br>68.5% |
| Number of plain-text files | 6,054<br>1.7% | 10,819<br>4.7% |
| Number of binary files | 126,689<br>34.7% | 60,953<br>26.7% |
| Entire websites without any META tags useful in search | 22<br>18.6% | 10<br>11.9% |
| Average number of META tags useful in search, per markup-language file | 3.06 | 0.89 |
| Average number of unique types of META tags useful in search, per website | 13.55 | 5.74 |

Table 1. Comparison of overall website measures between Illinois and Arizona state government webs.

## 2. "Find-It!" and the "State GILS Consortium"

Many states have met in annual conferences of an un-chartered consortium. These conferences began in connection with interest in the Government Information Locator System, and their working name reflects that. But, in subsequent years, the conferences have broadened their scope and now address many electronic information management issues of mutual interest (e.g., see the last two conference webpages at (GILS-4, 2002) and (GILS-5, 2003). In these conferences, and an associated e-mail list, state representatives and implementers share ideas, recommendations, lessons learned, open source or locally developed software, and software configuration materials such as search engine rule sets.

Several states use a variant of the jointly developed "Find-It!" search engine system philosophy. ISL operates the Illinois-customized version called "Find-It! Illinois" (ISL, 2001). Metadata, conformant with a state-specific variation on a jointly developed state government classification thesaurus, is used as a principal input to a locally configured and operated search engine. Decentralized metadata authoring reflects both staff and budgetary limitations of the archival agency, and a desire to enfranchise agencies concerning how their documents are represented to users. Multiple vendors offer search engines, which may differ somewhat in indexing and prioritization principles. For effective search to result, Find-It!

systems rely on high-quality metadata within the individual documents.

Find-It! search facilities reflects work to improve precision and recall in queries for state government information on the web. Because of the highly similar nature of many state agencies to their counterparts in other states, web-wide search engines generally exhibit considerably less precision in retrieval than might be hoped. The webpages of the counterpart agencies in other states are typically included in the search results, although a citizen is presumably only interested in the agency within his or her own state of residence. And, many web-wide search engines do not capitalize on classificatory information supplied inside webpages using META tags.

## 3. Statistical Observations

PEP began acquisition of the Illinois state government websites in mid-January 2002. To locate official Illinois state government websites, we employed a number of measures including (1) starting with a list previously developed by ISL, (2) performing deliberate breadth-first searching of know websites looking for hyperlinks to other seemingly official websites, (3) searching using popular Internet search engines and a number of general keywords, (4) searching using terms taken from the Illinois state government telephone directory, and (5) monitoring news outlets for announcements concerning new activities that might operate a website. Over thirty additional websites have been so discovered. State of Illinois websites have been found with host computer names ending in ".state.il.us", ".net", ".org", and ".com".

### 3.1 File types being used

Examining those files providing a file extension that indicates file type, some important results emerged. First, and completely unexpectedly, the comparative distribution of file type usage as a percentage of the total was within 1/2 of 1 percent for all file type categories. The distribution of file type usage is presented in table 2. Second, a preponderance of the file types utilized in both states was found to be markup-language files. Third, of the numerous word-processor-like file formats from which agencies might choose, these two states have almost exclusively adopted Adobe Portable Document Format (PDF). Fourth, the use of file formats specific to all other application programs is minimal. Note that many scripts used to serve document files do not report file type via file extension, but instead report it in the header used in the web (HTTP) transfer of the file. Such files are not included in the results in table 2.

| File Type | Illinois | Arizona |
|---|---|---|
| Markup language | 67.0 % | 67.4 % |
| Word processor documents | 16.4 % | 16.3 % |
| ASCII text documents | 4.5 % | 4.5 % |
| Still images | 11.0 % | 10.8 % |
| Other applications | 0.2 % | 0.1 % |
| Executable code | 0.7 % | 0.6 % |
| Other (unknown) | 0.2 % | 0.2 % |

Table 2. Comparison of file type distribution between Illinois and Arizona state government webs.

### 3.2 Metadata permeation within agency document inventories

Included within table 1 are two measures of the extent to which agencies in the two sampled states are currently providing metadata useful for the Find-It! search facilities. The state libraries of both states have been encouraging and facilitating the incorporation of metadata for over two years. However, the costs of retrofitting metadata into large government document inventories are not necessarily funded. Further, as

state libraries or archives are generally not a super-ordinate to other agencies in the government, the priority other agencies associate with retrofitting metadata may be low. Both states have several agencies which have not yet begun to add even one line of metadata. Other agencies, possibly those employing document management systems to aid them in inventory control of thousands of webpages, have tens of thousands of HTML META tags already embedded in documents. Both of these state libraries make available tools and instructional materials to facilitate the process of metadata generation and application, but they are not performing the classification work for the individual agencies.

The statistic on the average number of META tags useful in search, per markup-language file, reflects the elimination of META tags concerning the identity of the authoring tool or templates used in the creation of the file. While it would be possible for an authoring tool vendor to program their web browser software to behave differently based on information contained within META tags (e.g., through the imposition of a set of style-like backgrounds and borders), inspection of several Illinois agency websites did not find a case where such processing was in use. The "generator", "template", and "progid" META tags, plus Microsoft tags related to style and border defaults are ignored herein.

### 3.3 Website sizes and growth

A number of assumptions made in the development of the freeware subassemblies used in PEP combined to be problematic for the automatic population of an electronic archive. In particular, the widespread assumption that the characters following the identifier of a website host computer in a Uniform Resource Locator (URL) somehow equate to a physical location is problematic. Multiple problems were encountered where directories and files, named per character strings taken from URLs, resulted in erroneous operation of PEP components or standard UNIX commands. Differing metacharacters (characters with a special meaning to a particular host machine or program) between website host machines and the PEP host machine were also problematic. A very few websites had to be excluded from this analysis as their implementation technologies proved too dissimilar from the design basis of PEP components. Nevertheless, the statistics presented here represent our best efforts to date to detect and correct, or at least, to compensate for, this variety of errors.

The Illinois web grew much faster than expected in this period, particularly for its largest websites. The largest Illinois website in terms of bytes, the Illinois Pollution Control Board, increased in size by a factor of 50 in just over nine months. Growth rates of such magnitude can be highly problematic for both the website administrative staff and for the associated electronic archives. Other websites exhibited substantial reductions in size, generally followed by a return to something like their formal size, probably reflecting substantial redesign. Most websites exhibited continued, manageable growth.

It seems unlikely that the extreme rates of web authoring encountered for a very few agencies will be sustained. It seems more probable that they represent bursts of activity within the agencies, or capitalization upon an existing (“born digital”) electronic document collection that was easily reformatted. Considering the size of the state government staff in Illinois, it seems highly unlikely that hundreds of megabytes of documentary materials are being produced as a routine daily event. The spider-produced size of the website collection we process increased from 5.6 gigabytes in January 2002 to 27.7 gigabytes in September 2002. To support an increase of 22.1 gigabytes in roughly 35 weeks time, figuring 40 hours per week of uninterrupted ("no breaks") typing at 50 words per minute (where 5 bytes define one word), and that 72 percent of the current Illinois web size in bytes is made up of text files or markup-language files (which cannot contain embedded images that would contribute a disproportionate number of bytes) would require the employment of 758 such abused typists. It's clear from the occupancy of the office buildings in Springfield that these battalions of employees do not exist.

Exporting "born digital" (i.e., created using a word processor) documents into web formats such as HTML or PDF would be far easier than the initial typing and formatting. We might expect agencies to

exhaust their supply of born digital materials, with the eventual tapering off of the rate of increase in their websites. The Illinois Pollution Control Board website has exhibited very far less growth in size since mid July, possibly for this reason. Between August 14, 2002 and November 4, 2002, the version-controlled copy of this website increased in size by only 9.4 percent. Any number of other agencies may have sizeable inventories of born digital materials they will eventually post to the web, so growth rates of the electronic archives of an individual agency website should not be presumed to be regular and predictable.

## 4. Conclusions

Agency websites were observed to often be very volatile, especially in size. This volatility means government information is being lost now, and measures to prevent such loss cannot wait. The urgency for construction of suitable electronic document archives is as great as the importance of the government documents being lost. While not all government documents are high quality materials such as would have been archived in print form, web publishing should not be assumed to imply unimportance. Further, archival in electronic form can be a means of cost reduction for archives in that individual handling of documents over the life of the archive might be eliminated, or much reduced, by program-controlled processing of whole classes of documents simultaneously.

If an electronic document archive is to be constructed, very much the same metadata that supports search through the current contents of the web can also contribute to the search of the archive. However, measures must be taken to acquire and retain metadata for document types that do not support the embedding of metadata. And, metadata specific to identification of the version or dates of applicability of a document must be employed.

Metadata retrofitting has only begun, and may complete, if at all, only in the distant future. Alternatively, we need sufficiently capable information retrieval tools so as to be able to achieve reasonable usability of electronic document archives without dependence on embedded metadata. An automatic facility for analysis of whole-text materials and inference of metadata might suffice, but such a facility should be open, in order to interoperate with the various search engines in use. Further, such a system should scale economically, presumably utilizing parallel computing on a group of commodity workstations. The "brute force" computation hosts and proprietary software presently available appear to already be insufficient for the scale of government document collections.

Considering the expense of retrofitting metadata, it would be good to know that improved searching results. Formally controlled testing needs to address the cost/benefit ratio of a variety of ways through which web document collections can be made searchable. For example, formal cataloging by trained staff at a central facility (such as a state library) might be the best controlled, but also probably the most expensive option. But, is agency-authored metadata reliable, even when provided with a standardized thesaurus? This seems a reasonable question, considering communications problems in dealing with the distributed authoring and administration of numerous websites within government. Automated methods for topical metadata inference, by applying data mining and clustering techniques on the text of documents, is another possible lower-cost solution, if provably effective. Substantiated canceling of plans for metadata retrofitting could save government agencies much effort.

## 5. Future Work by the Illinois State Library

ISL is putting in place a two-pronged approach to the retention of electronic documents. Combined, they will form a consistent data warehouse, useful in data mining experimentation concerning synergistic combinations of agencies and in reducing overlap of expenditures. Sudden major changes such as the emergence of a Homeland Defense effort, or the transition of most of the Illinois executive branch can place urgent information demands on government staff. Data mining operations can inform concerning opportunities to leverage existing government information, and to identify synergistic combinations of

activities based on the literature produced by the groups involved. Such information can both speed response and reduce costs.

The current PEP spider-based duplication of whole agency websites is attractive for its relatively low cost of operation. But, this approach suffers from (1) cost increases as a function of web growth, (2) inability to identify webpages potentially most important to history, (3) many assumptions in component design that do not adequately deal with variances in web and script address spaces that can be encountered in URL-based access to documents, and (4) conflict with end-user expectations when file-based archival is used instead of document-based archival. We are seeking the resources to incorporate multiple technical and organizational lessons learned into a next-generation toolset that is specific to the construction of web archives that can be coherently browsed and searched by the citizenry. In addition to the identification of user communities and their information needs, effective and intuitive user interfaces incorporating time will be necessary to construct queries spanning multiple retained versions. Fiscal pressures on the states of the consortium provide strong incentive for the adoption of complete, open technical solutions to mutual problems in government electronic document archival.

A second effort will create a web-accessible electronic document depository facility where state agencies will deliberately deposit (upload) their materials that are designated to be permanently retained. By employing the agency staffs in deciding which documents warrant long-term retention, it is hoped the quantity of data being archived will be very substantially reduced. Also, much more complete metadata coverage of those documents should be practicable. This effort is underway now, under IMLS funding. Multiple data mining efforts will become possible once an Illinois data warehouse is assembled and sufficiently groomed (standardized).

Lastly, consortium states urgently need, and do not have, an archival and search technology option that is economical, open, and sufficiently scalable to deal with the volume of data typical of government agencies. Minimal, and extremely expensive search engine options are available commercially, but these do not expose their result-ranking algorithms to government configuration or control. Further, hardware implementations have not incorporated parallelism based on commodity computers, so high-capacity, single-node machines are prohibitively expensive for most government use. We are also seeking funding for an open, integrated archival and search system, leveraging author-provided metadata, but also capable of metadata inference based on data mining and information retrieval techniques. This system would be built from commodity workstations, with operations in parallel to economically support scaling system capacity to the needs of large agencies or large user populations.

## References

(GILS-4, 2002) Arizona State Library, Archives and Public Records, hosts. 4th State GILS conference, Scottsdale, AZ, April 24-27, 2002.
http://rpm.lib.az.us/4thGILS/index.html

(GILS-5, 2003) Illinois State Library, hosts. 5th State GILS conference, Lisle, IL, April 7-10, 2002.
http://www.cyberdriveillinois.com/library/isl/gils/gils_conf.html

(ISL, 2001) Illinois State Library. "Find-It! Illinois" homepage
http://finditillinois.org/

(Jackson, 2002) Larry S. Jackson. Preserving Electronic Publications project -- GSLIS materials webpage
http://www.isrl.uiuc.edu/pep/

(PEP, 2001) Joe Natale, Principal Investigator. Preserving Electronic Publications project homepage.
http://www.cyberdriveillinois.com/library/isl/lat/pep/pep.html